The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1950–1972
DOI: 10.1214/15-AOAS858

ANALYSIS OF MULTIVIEW LEGISLATIVE NETWORKS WITH STRUCTURED MATRIX FACTORIZATION: DOES TWITTER INFLUENCE TRANSLATE TO THE REAL WORLD?

By Shawn Mankad and George Michailidis

Cornell University and University of Michigan

The rise of social media platforms has fundamentally altered the public discourse by providing easy to use and ubiquitous forums for the exchange of ideas and opinions. Elected officials often use such platforms for communication with the broader public to disseminate information and engage with their constituencies and other public officials. In this work, we investigate whether Twitter conversations between legislators reveal their real-world position and influence by analyzing multiple Twitter networks that feature different types of link relations between the Members of Parliament (MPs) in the United Kingdom and an identical data set for politicians within Ireland. We develop and apply a matrix factorization technique that allows the analyst to emphasize nodes with contextual local network structures by specifying network statistics that guide the factorization solution. Leveraging only link relation data, we find that important politicians in Twitter networks are associated with real-world leadership positions, and that rankings from the proposed method are correlated with the number of future media headlines.

1. Introduction. There is a growing literature that attempts to understand and exploit social networking platforms for resource optimization and marketing, as it is a major interest for private enterprises and political campaigns attempting to propagate particular opinions or products [NYTimes (2011, 2012, 2013)]. An important problem is the identification of influential individuals that facilitate communication over the network. In this paper, we develop a modeling approach that captures influence from multiple networks that feature different link relations between the same set of nodes (e.g., Twitter accounts). Such multiview data are increasingly common due to the complex structure of many networking platforms. Specifically, we analyze three different types of networks that are commonly derived from Twitter data, each composed of either weighted or binary links.

Received October 2014; revised July 2015.
Key words and phrases. Matrix factorization, networks, influence, Twitter.

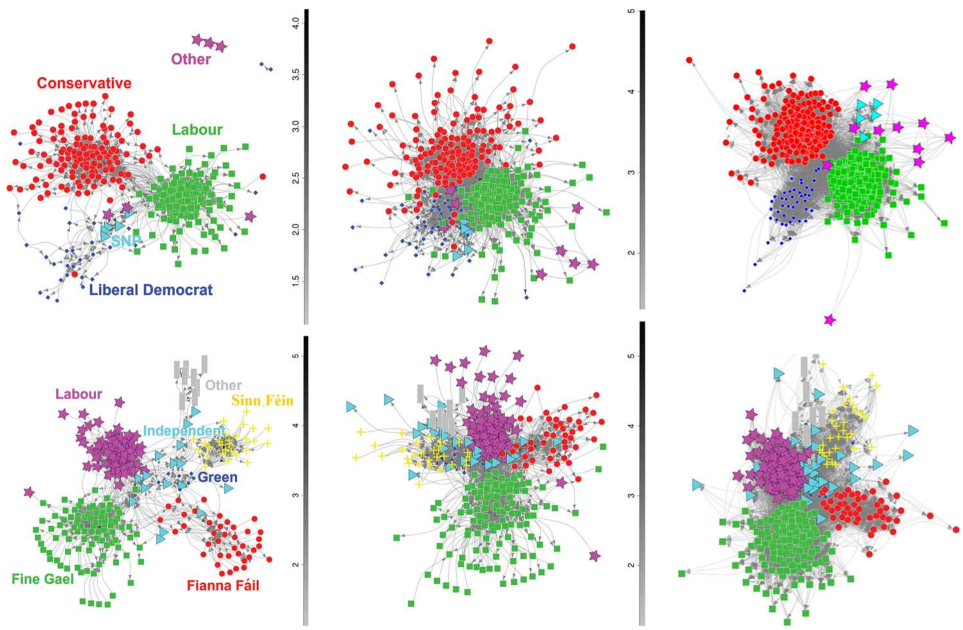
Twitter is a popular platform with over 270 million active accounts each month as of September 2014 [Twitter (2014)]. Twitter allows accounts to post short messages of 140 characters or less, commonly referred to as "tweets," that can be read by any visitor. A tweet that is a copy of another account's tweet is called a "retweet." Within a tweet, an account can mention another account by referring to their account name with the @ symbol as a prefix. Accounts also declare the other accounts they are interested in "following," which means the follower receives notification whenever a new tweet is posted by the followed account. These three directed actions define political Twitter networks that we analyze in this work.

The first network is a retweet network, where links are directed and weighted to denote the log-number of retweets from one account to another over an interval of time. The second network is also composed of directed and weighted links that denote the log-number of mentions one account gives another. The third network is constructed with directed binary links that denote the follower and followed relationships between accounts.

These three networks, each featuring 416 Members of Parliament (MPs) in the United Kingdom, are drawn in the top panel of Figure 1, where accounts are registered to 172 Conservative MPs, 185 Labour, 43 Liberal Democrats, 5 MPs representing the Scottish National Party (SNP), and 11 MPs belonging to other parties. There are 650 MPs forming the House of Commons, the lower house in the bicameral legislative body for the United Kingdom. Each MP is democratically elected to represent constituencies for five year terms, though often elections are held more frequently when Parliament is dissolved.

The second set of political Twitter networks that we analyze are drawn in the bottom panel of Figure 1. Each network is composed of 348 nodes that represent the accounts of Irish politicians and political organizations at all levels of government, including the President of the Republic of Ireland, members of the local and national government, and elected representatives for the European Union.

The raw data for both data sets, collected and processed by Greene and Cunningham (2013), consists of approximately 500,000 tweets and 40,000 follower links from late 2012. An empirical pattern observed in these data and also in previous studies [Huberman, Romero and Wu (2008)] is that the follower network is very dense in contrast to the retweet and mentions networks. Almost all politicians interact via retweeting or mentioning with a smaller number of other accounts, relative to their follower declarations. Moreover, users with many followers post updates less often than those with fewer followers [Huberman, Romero and Wu (2008)]. Such empirical findings suggest that not all links are created equally, and usually the follower network is discarded because it does not accurately capture patterns of conversation. However, each network, including the follower network, contains



(a) Retweet network (b) Mentions network (c) Follows network

FIG. 1. *The top panel shows networks of UK Members of Parliament and the bottom panel shows networks of Irish politicians and political organizations. Node color and vertex shapes denote party affiliation. The average degree for the UK Retweet, Mentions and Follows network is 9.13, 25.51 and 65.25, respectively. The average degree for the Irish Retweet, Mentions and Follows network shown in the bottom row is 5.81, 15.28 and 48.44, respectively.*

meaningful information, especially since we only consider the population of politicians in a specific legislative body instead of a broad set of users or even the entire Twitter userbase.

Previous research has found that Twitter and other social networking platforms help facilitate communication between politicians, government agencies and the broader public. Golbeck, Grimes and Rogers (2010) find by text mining tweets that members of the United States Congress employ Twitter for primarily two purposes: information dissemination and self promotion. Tumasjan et al. (2010) find that the number of tweets from the general public mentioning a political party or politician is a valid indicator of political sentiment and a good predictor of federal election results in Germany. More recently, similar results have been found for federal elections in Australia and the U.S. House of Representatives [Unankard et al. (2014), McKelvey, DiGrazia and Rojas (2014)]. In contrast to these previous works, we rely only on the link relations, so-called "meta-data," among politicians to measure influence and identify conversation flows with network analysis.

Approaches that utilize content analysis can face significant challenges associated with text and image analysis (accounts can post a photo within a tweet), such as language differences, tone and sentiment characterization, and so on.

There has been extensive work on ranking nodes on a network by their importance primarily motivated by search on the World Wide Web. We find our proposed method compares favorably for ranking politicians against two seminal works called PageRank [Page et al. (1999)] and HITS [Hyperlink-Induced Topic Search; Kleinberg (1999)]. The idea behind PageRank is to use as a measure of importance an estimate of the probability of reaching a given node by randomly following edges. HITS utilizes the so-called authority and hub scores, which are computed by the leading eigenvector of A^TA and AA^T, respectively, where A is an adjacency matrix.

Our main goal of identifying influential politicians is also closely related to role identification, which aims to assign roles based on local connectivity patterns. Typically, role analysis methods rely on analyzing ego networks (the union of a node and its neighbors), network statistics or graph-coloring techniques [Salter-Townshend and Murphy (2015)]. Also note that while there have been many recent advances in community detection, including the stochastic block model, latent position cluster models and others [see Fienberg (2012), Salter-Townshend et al. (2012) for survey articles], the task in this article is different from typical community detection, which aims to extract groups of nodes that feature relatively dense within-group connectivity and sparser between-group connectivity. That said, community detection could help guide a search for influential politicians. For instance, an analyst may examine each network separately by first discovering communities, if unknown, then searching for interesting network statistic profiles within each group. There are in principle many ways to combine community detection with network statistics for the identification of influential nodes, (e.g., politicians), but it remains unclear which is the preferred method. In this paper, we integrate both steps together to address this issue. The proposed factorization model is also able to emphasize nodes with interesting path-related properties by incorporating node-level statistics that capture these nonlinear relationships, thus leading to more interpretable measures of influence and substructure.

The main idea is to guide the mapping of the multiview networks to lower-dimensional spaces using structured matrix factorization. Nonnegativity constraints are also imposed on the lower-dimensional spaces to improve data representation and structural discovery. Such constraints have been popularized with the nonnegative matrix factorization (NMF) and Semi-NMF, where one or all matrix factors are composed of only nonnegative entries and have been shown to be advantageous for data representation [Lee and Seung (1999), Ding, Li and Jordan (2010)]. As validation, we find

that important politicians identified using our modeling approach are associated with real-world leadership positions, and that rankings from the proposed method are significantly correlated with future media headlines. The consistent findings between both data sets suggest the model can be a relatively straightforward technique for identifying influential individuals with political Twitter networks from other countries that feature different government structures, and that it can complement the potentially more involved content analysis for related tasks.

The next section introduces the matrix factorization model, followed by estimation details in Section 3. Section 4 summarizes and compares results of the proposed model against alternative methodologies with UK MPs and Irish politicians. This article closes with a brief discussion in Section 5.

2. Structured semi-NMF for influence discovery. The use of low-rank approximations to network related matrices follows a long line of previous work. In classical spectral layout, the coordinates of each node are given by the Singular Value Decomposition (SVD) of the Laplacian matrix [Koren (2005), Brandes, Fleischer and Puppe (2006)]. Recently, there has been extensive interest in spectral clustering [Rohe and Yu (2012), Rohe, Chatterjee and Yu (2011)], which discovers community structure in the eigenvectors of the Laplacian matrix.

Low-rank approximations satisfying different constraints other than orthonormality are also popular. For instance, NMF has been proposed for overlapping community detection on static [Psorakis et al. (2011), Wang et al. (2011)] and dynamic [Lin et al. (2008)] networks. When overlaps among communities exist, an advantage of NMF over spectral clustering is that NMF can still find basis vectors for each community, while orthogonality of SVD makes it unlikely that the singular vectors will correspond to each of the communities [Xu, Liu and Gong (2003)]. The basic framework for NMF in network analysis is $A \approx UV^T$, where A is an adjacency matrix and $U, V \in \mathbb{R}_{\geq 0}^{n \times K}$. Written in element form,

$$A_{ij} \approx U_{i1}V_{j1} + \cdots + U_{iK}V_{jK},$$

one can easily see that each edge of the given network is approximated with a nonnegative sum. Consequently, each term in the sum, $U_{ik}V_{jk}$, represents the contribution of the kth latent structure (often capturing community structure especially when decomposing sparse adjacency matrices [Mankad and Michailidis (2013b)]) to the edge from i to j. Edge decompositions can be aggregated by node or one can use the rows of V to directly determine node community membership. The factors are found by minimizing

$$\min_{U \geq 0, V \geq 0} \|A - UV^T\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The optimization can be performed using gradient-descent algorithms for penalized optimization. Given that the proposed model in this article utilizes nonnegativity, we follow a similar algorithmic approach to the NMF literature.

Enforcing nonnegativity on a single matrix factor was first proposed in Ding, Li and Jordan (2010) with the so-called Semi-NMF to improve interpretability of the resultant factorizations with data of mixed signs. We utilize the flexibility of Semi-NMF and extend it to the network setting with a structured approach that incorporates graph geometry into the factorization through user-specified matrices. In particular, we aim to utilize the many node-level statistics that have been proposed in the network literature to guide the factorization solution. Next we introduce the model for singleview networks, then extend to multiview networks, followed by estimation procedures in the next section.

2.1. *Singleview networks.* Let A denote the adjacency matrix from a single, given network with n nodes. We start with the following graph Structured Semi-NMF model of Mankad and Michailidis (2013a):

$$\min_{\Lambda,\Theta\geq 0}\|A - S\Lambda\Theta^T\|_F^2, \tag{1}$$

where $S \in \mathbb{R}^{n\times D}, \Lambda \in \mathbb{R}^{D\times K}$, and $\Theta \in \mathbb{R}^{n\times K}_{\geq 0}$. Note that Θ is nonnegatively constrained, but Λ is not, which is why the model fits into the Semi-NMF framework. Each factor in the product $\Lambda\Theta^T$ is estimated from the data and provides coefficients for each node that represent the given adjacency matrix in terms of S.

The S matrix is composed of D node-level statistics that are specified by the analyst before performing the factorization to emphasize nodes that drive influence. There is an extensive literature in network analysis providing potential node-level statistics [Newman (2010)]. In our analysis, the S matrix is constructed using $D = 4$ network statistics and has form

$$S_i = [\text{clustering coefficient}_i, \text{betweenness}_i, \text{closeness}_i, \text{degree}_i],$$

where $i = 1, \ldots, n$. The *clustering coefficient* for a given node quantifies how close its neighbors are to forming a complete graph [Newman (2010)]. A higher clustering coefficient will emphasize politicians that "create buzz." *Betweenness* [Freeman (1979)] and *closeness* [Newman (2010)] rely on shortest path statistics and capture important links from hub nodes. *Degree*, the number of connections a node has obtained, ensures that active politicians within communities are emphasized in the factorization.

If there are no node-specific values that are obvious to use for S, one can start with many candidate node-level statistics and search for subsets that fit the data well while maintaining interpretability. This strategy will be

discussed further below to also show robustness and assess the specification of S in our application. Instead of searching over node-specific statistics, one could also be tempted to set $S = I_{n\times n}$ to be the identity matrix. In this case, the factorization is essentially the standard Semi-NMF factorization. Our results show that the Semi-NMF model performs similarly to classical importance measures, like PageRank and HITS, which should be preferred due to their more efficient implementations.

The proposed model implies certain connectivity dynamics that can be seen when equation (1) is written in element form

$$A_{ij} \approx (S\Lambda)_{i1}\Theta_{j1} + \cdots + (S\Lambda)_{iK}\Theta_{jK},$$
$$(S\Lambda)_{ik} = S_{i1}\Lambda_{1k} + \cdots + S_{iD}\Lambda_{Dk}.$$

For any node i, outgoing edges are controlled by its local topological characteristics, as measured in S, and how communities load onto the statistics in S, given in the columns of Λ. When multiplied together, $S\Lambda$ form centroids in a K-dimensional space that capture the outgoing node influence from each of the communities. The receiving node j in an edge is determined by the jth row of Θ, where larger values mean the node is more likely to have incoming connections and, hence, greater influence.

Due to nonnegativity and the fact that Θ modulates incoming connections, we accomplish our ultimate goal of measuring overall influence for the ith node by taking its cumulative sum of importance to each community

$$\mathcal{I}_i = \sum_{k=1}^{K} \Theta_{ik}. \tag{2}$$

As illustrated in the supplemental article [Mankad and Michailidis (2015)] on a toy example, the S matrix plays a pivotal role in the factorization, and causes $\mathcal{I}$ to be an effective importance measure even with its relatively simple definition.

Next we propose an extension of this model to the multiview setting found in political Twitter networks.

2.2. *Multiview networks.* Let A_m denote the adjacency matrix from the corresponding Twitter network, where $m = \{\text{retweet}, \text{mentions}, \text{follows}\}$. We extend the singleview model with

$$\min_{\Lambda_m, \Theta\geq 0, V_m \geq 0} \sum_m \|A_m - S_m\Lambda_m(\Theta + V_m)^T\|_F^2, \tag{3}$$

where $S_m \in \mathbb{R}^{n\times D}, \Lambda_m \in \mathbb{R}^{D\times K}$, and $\Theta, V_m \in \mathbb{R}^{n\times K}_{\geq 0}$. Θ is common to all m networks to capture general structure and makes the objective function nonseparable, whereas V_m reveals network-specific structure and also implicitly weights each network according to its importance in the factorization.

The S_m matrices are defined similarly to the singleview case, using node-level network statistics. We define S_m using the same four network statistics for each network view. Weighted versions of the clustering coefficient and degree are utilized for the Retweet and Mention networks in order to take into account the frequency of interaction between politicians, since the frequency should help measure the strength of a relationship [Barrat et al. (2004)]. For instance, a weighted network statistic will distinguish between a politician that is retweeted by the same account hundreds of times versus retweeted once. The model does allow for different statistics to be defined with each network view, which may be advantageous in other contexts.

The final importance measure $\mathcal{I}$ can also be calculated similarly using equation (2). Since Θ is common to all networks, the importance measure is a result of integrating multiple network views in addition to structured discovery.

3. Algorithms. The estimation algorithm we present is an iterative one that cycles between optimizing with respect to Θ, V_m and Λ_m with the following updates:

$$\Theta = \sum_m A_m^T S_m \Lambda_m (\Lambda_m^T S_m^T S_m \Lambda_m)^{-1},$$

$$V_m = A_m^T S_m \Lambda_m (\Lambda_m^T S_m^T S_m \Lambda_m)^{-1},$$

$$\Lambda_m = (S_m^T S_m)^{-1} S_m^T A_m (\Theta + V_m)((\Theta + V_m)^T (\Theta + V_m))^{-1}.$$

The updates are based on alternating least squares (ALS) and derived through standard arguments [Kroonenberg and de Leeuw (1980)], which are shown in the supplemental article [Mankad and Michailidis (2015)].

Technically, both Θ and V_m require solving nonnegatively constrained least squares problems, which result in high iteration costs. So, instead of exactly solving the constrained least squares problem, we follow a heuristic that solves for an unconstrained solution, then sets any entry less than a user-specified constant to that constant. Projecting to a small constant instead of zero follows the discussion in Gillis and Glineur (2008) and Katayama, Takahashi and Takeuchi (2013) to overcome numerical instabilities that occur when too many elements are exactly zero.

Theoretical properties are difficult to obtain due to the projection step. Yet this approximation is computationally efficient, easy to implement, and has been shown to achieve high quality solutions [Berry et al. (2007)]. The algorithm easily scales to networks with tens of thousands of nodes. For even larger networks on the order of millions of nodes, low-rank factorizations should be found using recent algorithmic advances that exploit parallel computing architecture [Gemulla et al. (2011), Recht and Ré (2013)]. For

our data, we find that the alternative least squares algorithm is straightforward to implement and able to recover meaningful factorizations in a timely fashion.

In the supplemental article [Mankad and Michailidis (2015)], we also discuss an alternative updating approach for Θ and V_m that is similar to the popular "multiplicative updating" for NMF. While this approach is also very easy to implement, we find the ALS algorithm more numerically stable in higher dimensions.

3.1. *Initialization and convergence criteria.* An advantage of the ALS algorithm is that only Λ_m needs to be initialized if the order of the updates is Θ, V_m, Λ_m. Moreover, recall that Λ_m is unconstrained, thus bypassing the difficulties of initializing the nonnegative factors which have received extensive focus in the NMF literature. We find stable results by initializing Λ_m with normally distributed entries having unit mean and variance.

Another important issue is specifying the rank of the matrices Θ and V_m. Ideally, the rank should be equal to the number of underlying communities and can be ascertained by examining the accuracy of the reconstruction as a function of rank. In principle, one could also apply cross-validation procedures for matrix factorization [Owen and Perry (2009)], though this may become cumbersome with sparse or extremely large-sized networks.

We follow a strategy similar to using a scree plot to choose the number of components to retain in Principal Component Analysis [Jolliffe (1986)]. To our knowledge, this rank selection approach has not been previously pursued in the context of NMF or Semi-NMF. Shown in Figure 2, we find that ranks greater than six (roughly the number of underlying political parties) yield little marginal explanatory power. Each subfigure is constructed by plotting the best fitting factorization over all possible network statistic subsets of size two through four. The appropriate rank of the matrices Θ and V_m is stable across the S_m subsets, though there appears to be significant improvement when S_m is defined with at least three of the network statistics. We keep all four network statistics when defining S_m for our analysis.

Last, we discuss convergence criteria used for the ALS algorithm. Let $\mathcal{O}^{(i)}$ denote the value of the objective function at iteration i. Then the algorithm stops when $\frac{|\mathcal{O}^{(i)}-\mathcal{O}^{(i-1)}|}{\mathcal{O}^{(i-1)}} \leq \varepsilon = 10^{-4}$. We find in all our investigations that the algorithm converges within 50 iterations. $\varepsilon = 10^{-4}$ is also used for the projection threshold.

4. Analysis of the political multiview Twitter networks.

4.1. *Does Twitter influence translate to the real world?* Using the best rank six factorization with S_m defined with all four network statistics, we

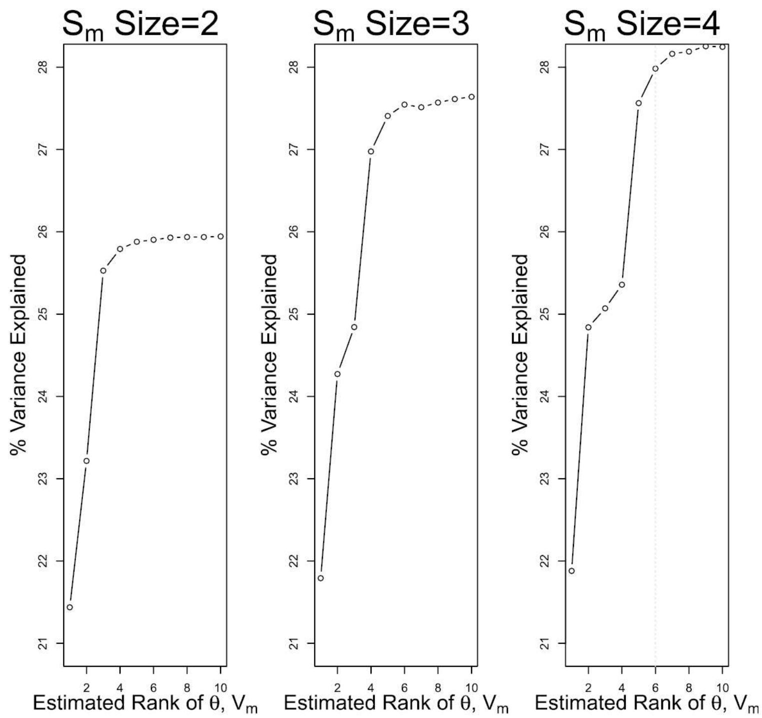


FIG. 2. *The percentage of variance explained [$100*(1-\sum_m \|A_m - \hat{A}_m\|_F^2/\|A_m - \hat{\mu}_m\|_F^2)$, where $\hat{\mu}$ is a matrix filled with the average value of A_m] for the Structured Semi-NMF with different constructions of S_m. Plotted is the most accurate model over thirty trials with random initializations for Λ_m at each possible specification. We use the best rank six model with four network statistics composing S_m for the final analysis.*

rank MPs according to the estimated Θ and the importance measure defined in equation (2).

Figure 3 shows the importance scores from the Structured Semi-NMF, Semi-NMF, PageRank and HITS. PageRank and HITS are computed using the retweet network, which has been shown to capture conversation dynamics better than the other network types [Cha et al. (2010)]. Not surprisingly, the different importance measures are all positively correlated.

Accordingly, as shown in Table 1, there is general agreement between Structured Semi-NMF, Semi-NMF and HITS in the top ten important MPs. Many of these MPs held leadership positions in the coalition or Opposition cabinets. For instance, *Ed Miliband*, leader of the Labour Party and of the Opposition at the time of writing, is prominently emphasized in all rankings. *Tom Watson* was the Deputy Chair of the Labour Party, and *Chuka Umunna* is the Shadow Secretary of State for Business, Innovation and Skills. The exceptions are *Rachel Reeves*, who became the Shadow Secretary of State for Work and Pensions for the Opposition after the data was collected, and

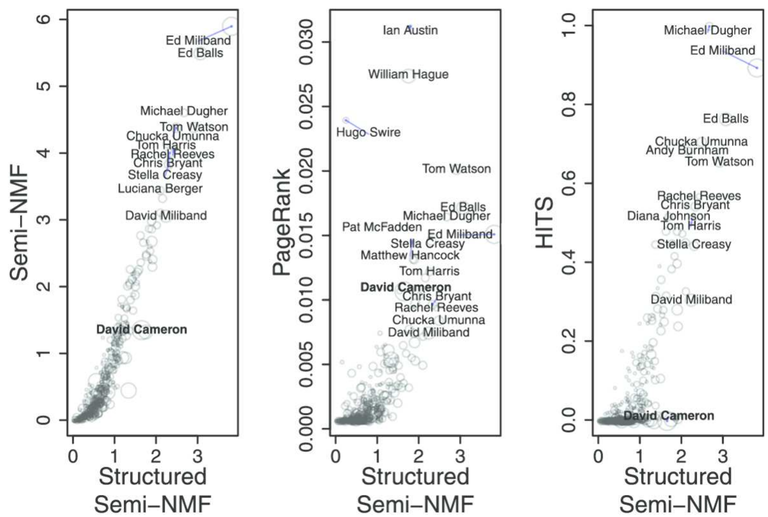


FIG. 3. *Importance scores based on Structured Semi-NMF, Semi-NMF ($S_m = I_{n\times n}$), PageRank and HITS (Authority Scores). PageRank and HITS are both calculated using the Retweet network, while the other measures utilize all three networks. The radius of the circle indicates the count of future newspaper headlines as measured with Lexis–Nexis. The top ten MPs for the methods in each scatterplot are labeled. David Cameron, who is Prime Minister and in boldface, was not in the top ten for any method.*

David Miliband, who held several important positions in previous terms prior to data collection.

Another commonality is that, with the exception of PageRank, every MP in the top ten is from the Labour Party. Labour MPs tend to be estimated as most important, followed by Conservative, and then Liberal Democrat MPs. The relative ranking among parties is consistent with the data, where Labour MPs tend to be the most active users in our data. Of the top fifty Twitter accounts in terms of number of retweets or mentions, only four are affiliated with another party—the Conservatives. The Liberal Democrats are even less active, ranked in the hundreds in terms of number of retweets or mentions. For instance, *Nick Clegg*, leader of the Liberal Democrats and Deputy Prime Minister at the time of writing, is typically the top-ranked member of his party at forty-nine with Structured Semi-NMF, forty with PageRank, and outside the top hundred with both Semi-NMF and HITS.

Activity in the data set is likely associated with longevity on Twitter. For instance, *David Cameron*, Prime Minister and leader of the Conservatives, is ranked twenty-nine with Structured Semi-NMF, sixty-eight with Semi-NMF, sixteen with PageRank, and two hundred and forty-two with HITS. Cameron joined Twitter just as the data was collected in October 2012, and, thus, may have artificially low levels of activity when compared against more recent data. In spite of these challenges, PageRank and Structured

TABLE 1

MP rankings and in parentheses the party and frequency that the MP appears in future headlines for Structured Semi-NMF, Semi-NMF ($S_m = I_{n\times n}$), PageRank and HITS (Authority Scores). L denotes Labour, C denotes Conservative

Rank	**Structured Semi-NMF**	**Semi-NMF**	**PageRank**	**HITS**
1	Ed Miliband (L, 2478)	Ed Miliband (L, 2478)	Ian Austin (L, 3)	Michael Dugher (L, 120)
2	Ed Balls (L, 580)	Ed Balls (L, 580)	William Hague (C, 771)	Ed Miliband (L, 2478)
3	Tom Watson (L, 253)	Michael Dugher (L, 120)	Hugo Swire (C, 57)	Ed Balls (L, 580)
4	Michael Dugher (L, 120)	Tom Watson (L, 253)	Tom Watson (L, 253)	Chuka Umunna (L, 203)
5	Chuka Umunna (L, 203)	Chuka Umunna (L, 203)	Ed Balls (L, 580)	Andy Burnham (L, 125)
6	Rachel Reeves (L, 54)	Rachel Reeves (L, 54)	Michael Dugher (L, 120)	Tom Watson (L, 253)
7	Stella Creasy (L, 178)	Chris Bryant (L, 164)	Pat McFadden (L, 1)	Rachel Reeves (L, 54)
8	Chris Bryant (L, 164)	Stella Creasy (L, 178)	Ed Miliband (L, 2478)	Chris Bryant (L, 164)
9	Tom Harris (L, 113)	Luciana Berger (L, 133)	Stella Ceasy (L, 178)	Diana Johnson (L, 105)
10	David Miliband (L, 489)	Andy Burnham (L, 125)	Matthew Hancock (C, 32)	Tom Harris (L, 113)

Semi-NMF with use of the S_m matrix are able to boost these key MPs importance, even though they interact via Twitter with their MP colleagues relatively infrequently.

We have so far seen anecdotal evidence that many MPs in leadership positions are emphasized by the different techniques. Next, we test in a regression setting whether these different measures of Twitter importance predict media coverage, which is measured using Lexis–Nexis (www.lexisnexis.com) searches of the number of times an MP's name appears in headlines from January 1, 2013, to October 17, 2013. This interval of time is strictly after the Twitter data was collected to avoid endogeneity issues. Because the headline counts were overdispersed, we use a quasi-Poisson regression. The mean and variance of the regression has form

$$\mathbb{E}(\text{HeadlineCount}_i) = \exp(\alpha + \beta \mathcal{I}_i + \gamma \,\text{Controls}_i), \tag{4}$$

$$\text{Var}(\text{HeadlineCount}_i) = \rho \mathbb{E}(\text{HeadlineCount}_i), \tag{5}$$

where $\rho \geq 1$ is estimated from the data. HeadlineCount is the headline occurrence frequency, $\mathcal{I}$ is derived using the different importance measurement techniques, and Controls contain the variables Age, Gender, Constituency Size, Political Party and an indicator variable denoting whether each MP represents a constituency within the city of London. Age is an important control variable, since we expect younger MPs to be more savvy with social media, which could affect their headline coverage. Similarly, we expect MPs with larger constituencies, certain political affiliations or London-based MPs to receive more media attention.

Additional discussion in the supplemental article [Mankad and Michailidis (2015)] shows the Poisson distributional assumption appears more valid when compared to other distributions for overdispersion, like negative binomial. Moreover, the quasi-Poisson results featured the smallest root mean squared error (RMSE) for all specifications that we discuss next.

In Figure 4, we examine the RMSE of the model when using only control variables, as well as control variables with each influence measure separately. We find that the model using the proposed factorization features the lowest RMSE, especially after removing an outlier, David Cameron, who received many more future headlines than predicted. As mentioned above, David Cameron joined Twitter just as the original data set was collected, potentially creating an artificially low presence on Twitter.

Table 1 in the supplemental article [Mankad and Michailidis (2015)] shows the full results for the estimated model with Structured Semi-NMF, where the corresponding coefficient is statistically significant and positive as expected. Specifying S_m leads to an importance measure that is associated with future media headlines even when controlling for other influence measures and demographic information, thus illustrating the importance of guiding the factorization solution.

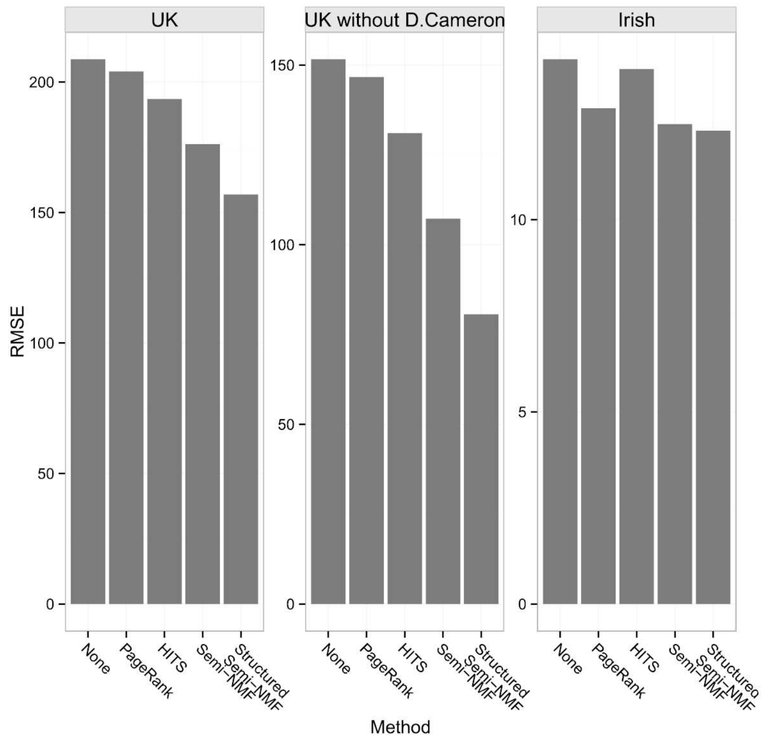
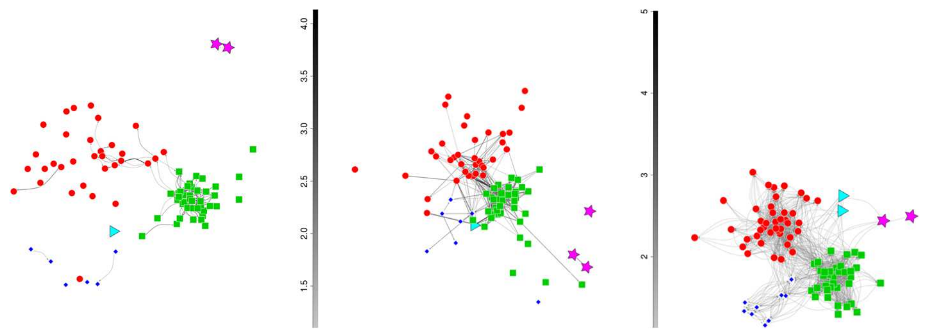


FIG. 4. *Root mean squared errors for the predicted number of headlines using different specifications of the regression model in equations (4) and (5). "None" refers to including only control variables. "PageRank" refers to the control variables plus the PageRank influence measure, "HITS" refers to the control variables plus the HITS influence measure, and so on.*

4.2. *Identifying important conversation flows.* Another advantage of the proposed factorization is that it can also be used to extract potentially important conversation flows. We construct subgraphs by keeping nodes in the top qth percentile of $\sum_k(\Theta + V_m)_{ik}$ to recover structure specific to each network view.

The Structured Semi-NMF does not incorporate party affiliation for the factorization. Yet it results in more interpretable subgraphs than the alternative approach in Figure 5 of looking at high degree nodes within each party. Shown in Figure 6, there are denser within and between party connections, and fewer isolated nodes. Moreover, with the exception of a handful of MPs, each node can reach every other node on the graphs. Thus, these networks help explain the influence rankings from the previous section by identifying paths through which interesting content flowed.

Tracing the flow of conversations in the 95 percentile subgraphs in Figure 7, we see that the Labour politicians tend to retweet each other of-



(a) Retweet network (b) Mentions network (c) Follows network

Fig. 5. *Subnetworks of UK Members of Parliament chosen by taking the highest degree MPs in each party, with color and vertex shapes denoting party affiliation. MPs are drawn in the same position as in Figure 1.*

ten. Many of the Labour MPs, including *Stella Creasy*, *Ed Miliband*, *Chuka Umunna*, *Rachel Reeves*, *Tom Watson* and others, were universally ranked as important in the previous section. *Ed Balls* from Labour interacts directly with *Greg Hands* of the Conservative party, who in turn forms a much

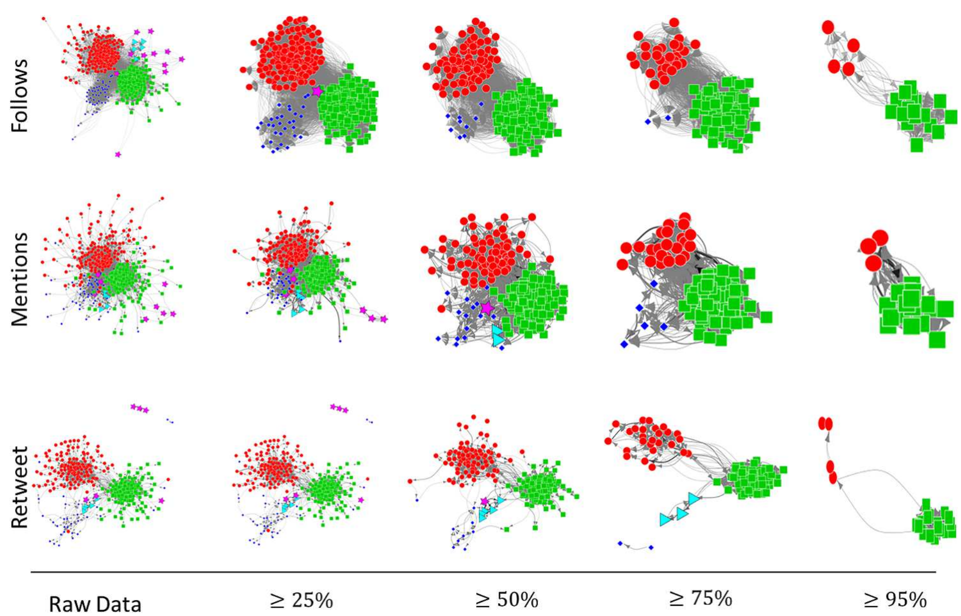


Fig. 6. *Networks of UK Members of Parliament, with color and vertex shapes denoting party affiliation. MPs in the top* q*th percentile of* $\sum_k (\Theta + V_m)_{ik}$ *are kept and drawn in the same position as in Figure 1.*

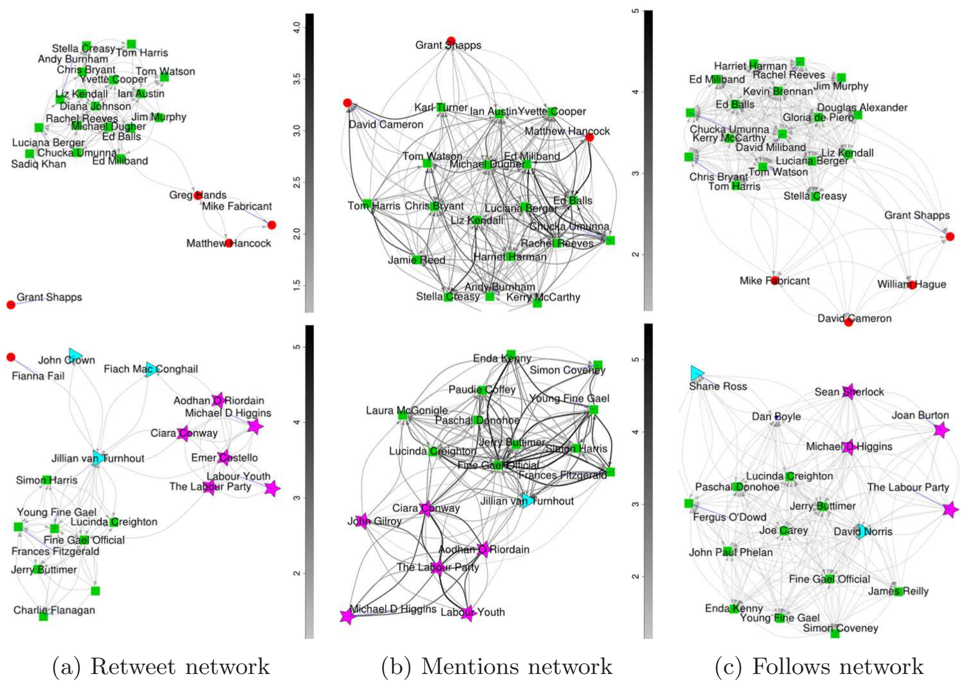


Fig. 7. *Subgraphs constructed for the UK MPs (top panel) and Irish politicans (bottom panel), whose nodes are in the top $q = 95$ percentile of $\sum_k(\Theta + V_m)_{ik}$. Graphs are redrawn to optimize vertex labels.*

smaller retweet clique with fellow Conservatives *Matthew Hancock* and *Mike Fabricant*.

Since retweeting can amount to an endorsement, while mentioning allows the author to control the content and sentiment, there are a greater number of cross-party mentions edges. For instance, *David Cameron* is mentioned often and followed by Labour MPs, elevating his importance on those specific networks, but is never retweeted. This illustrates the value of utilizing all three types of networks for measuring importance.

4.3. *Analysis of Twitter networks from the Irish political sphere.* We produce comparable, though less pronounced results with similar Twitter network data from the Irish political scene from late 2012. We organize the raw data again provided in Greene and Cunningham (2013) into the same three Twitter networks, each containing 348 nodes that represent the accounts of Irish politicians and political organizations. The data contains politicians from all levels of government, including the President of the Republic of Ireland, members of the local and national government, and elected representatives for the European Union.

A majority of accounts belong to members of the Irish national parliament, which is also a bicameral legislative body with elections held at least once every five years using a system [Coakley and Gallagher (2005)]. The lower house (Dáil Éireann) is the principal house in the Irish system and contains 166 elected members, the senate (Seanad Éireann) contains a mixture of 60 appointed and elected members. There are multiple political parties in the data: 33 Fianna Fáil, 127 Fine Gael, 6 Green, 20 Independent, 68 Labour, 22 Sinn Féin and 8 Others. Approximately 60 Twitter accounts are registered to political parties, for example, "Fine Gael Official," "Labour Women," etc.

After specifying S_m as before and setting $K = 7$ (chosen in a similar fashion), we plot the importance scores in Figure 8 and list the top ten accounts in Table 2 from the Structured Semi-NMF, Semi-NMF, PageRank and HITS. In contrast to the British MP dynamics, political organizations seem to play a much more important role in online conversations within the Irish political sphere, as there is broad agreement among the different importance measures that party organization accounts are highly ranked, such as *Fine Gael Official*, *Young Fine Gael*, and *The Labour Party*. Some politicians are also universally ranked as important. *Michael D Higgins*, the President at the time of writing, is ranked eleventh under the Structured Semi-NMF, thirteenth under PageRank and in the top ten for all other methods.

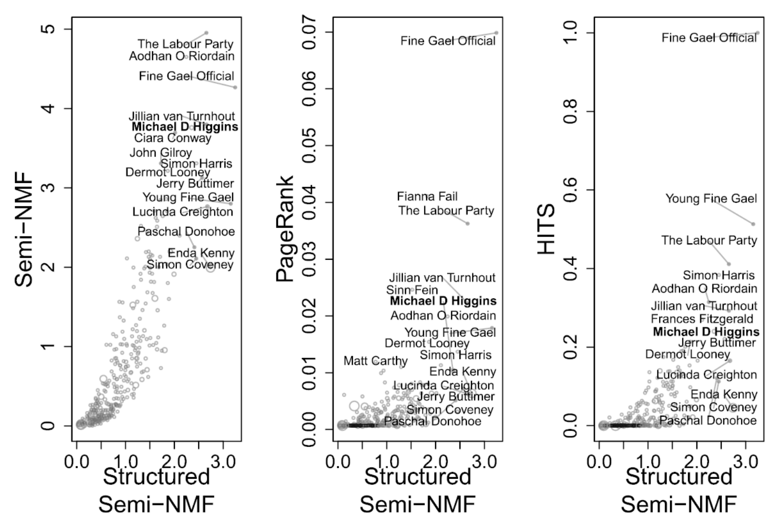


FIG. 8. *Importance scores based on Structured Semi-NMF, Semi-NMF ($S_m = I_{n\times n}$), PageRank and HITS (Authority Scores) are both calculated using the Retweet network. The radius of the circle indicates count of future newspaper headlines as measured with Lexis–Nexis. The top ten Irish politicians for the methods in each scatterplot are labeled. Michael Higgins, President, is boldfaced.*

TABLE 2

Irish politician rankings and in parentheses the party and frequency that the politician appears in future headlines for Structured Semi-NMF, Semi-NMF ($S_m = I_{n\times n}$), PageRank and HITS (Authority Scores). L denotes Labour, FG denotes Fine Gael, Ind denotes Independent and SF denotes Sinn Féin. There are no parenthetical headline counts or party names for political organizations

Rank	**Structured Semi-NMF**	**Semi-NMF**	**PageRank**	**HITS**
1	Fine Gael Official	The Labour Party	Fine Gael Official	Fine Gael Official
2	Young Fine Gael	Aodhán Ó Ríordáin (L, 1)	Fianna Fáil	Young Fine Gael
3	Enda Kenny (FG, 166)	Fine Gael Official	The Labour Party	The Labour Party
4	Lucinda Creighton (FG, 20)	Jillian van Turnhout (Ind, 0)	Sinn Féin	Simon Harris (FG, 4)
5	Jillian van Turnhout (Ind, 0)	Michael D Higgins (L, 25)	Jillian van Turnhout (Ind, 0)	Aodhán Ó Ríordáin (L, 1)
6	The Labour Party	Ciara Conway (L, 0)	Aodhán Ó Ríordáin (L, 1)	Jillian van Turnhout (Ind, 0)
7	Jerry Buttimer (FG, 2)	Simon Harris (FG, 4)	Young Fine Gael	Frances Fitzgerald (FG, 7)
8	Simon Harris (FG, 4)	John Gilroy (L, 3)	Dermot Looney (Ind, 0)	Michael D Higgins (L, 25)
9	Simon Coveney (FG, 10)	Dermot Looney (Ind, 0)	Simon Harris (FG, 4)	Jerry Buttimer (FG, 2)
10	Paschal Donohoe (FG, 4)	Jerry Buttimer (FG, 2)	Matt Carthy (SF, 0)	Dermot Looney (Ind, 0)

ods. *Jillian van Turnhout* is an appointed member of the Seanad Éireann and is consistently ranked highly by the different influence measures. Likewise, *Jerry Buttimer* is a member of the Dáil Éireann and formerly of the Seanad Éireann, and *Simon Harris* was elected to the Dáil Éireann in 2011 as its youngest member.

There are key differences, however, among the various importance measures. *Dermot Looney* is ranked in the top ten for Semi-NMF, PageRank and HITS, but nineteenth under Structured Semi-NMF. He seems to be ranked higher than one may expect, since Looney was part of a local government and served as mayor of the South Dublin County Council. *Lucinda Creighton* is ranked fourth for the Structured Semi-NMF, but is not in the top ten for other importance measures. At the time of data collection, Creighton served as Minister for European Affairs representing Ireland in negotiations on Ireland's EU/IMF bailout and the hosting of Ireland's presidency of the European Union. We also see that *Enda Kenny*, an Irish Fine Gael politician who has been the Taoiseach (prime minister) since March 2011, is ranked in the top ten only under the Structured Semi-NMF approach. He is ranked fortieth with Semi-NMF, thirty-fourth with PageRank and seventy-second with HITS.

The larger differences between the Structured Semi-NMF and other importance measures when compared to the UK MP results can be explained by the sparser input networks, as shown in Figure 9, which increase the effect of the S_m matrices. Figure 7 shows the conversation dynamics that help explain why certain accounts are ranked highly with the structured approach. For instance, we see that *Jillian van Turnhout*, an Independent, tends to be retweeted or mentioned by Fianna Fáil organizations in addition to Fine Gael, Labour and other Independent politicians. Accounts within the Labour party also form their own clique, centered around *Michael D Higgins* and the official Labour party account.

Finally, we test whether these different measures of Twitter importance predict media coverage with the same quasi-Poisson model as in equations (4) and (5). Headline occurrence frequency from January 1, 2013, to October 17, 2013, is again measured using Lexis–Nexis searches, $\mathcal{I}$ is derived using the different importance measurement techniques, and Controls contains the variables Age, Gender, Politician Type (local, presidential, Dáil Éireann, Seanad Éireann, European Union), Constituency and Political Party. Since the data contains politicians in local government, where, for example, exact constituency size is not easily defined for council members, we include a fixed effect for every unique electoral district or area. The 134 unique areas are identified using a number of online sources, including official party and candidate websites, newspaper articles and election results posted on https://electionsireland.org/. Party organization accounts are removed when estimating the regression model.

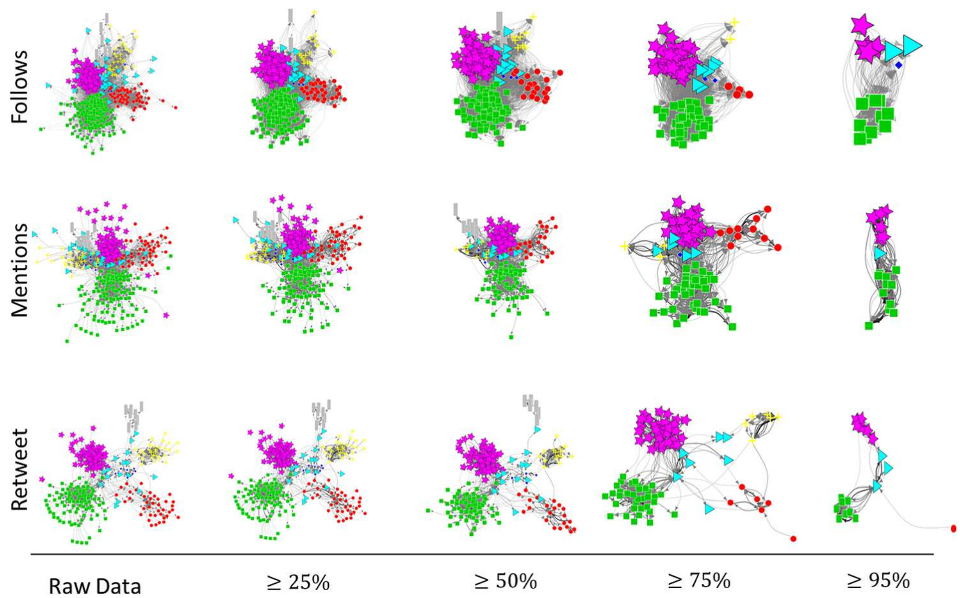


FIG. 9. *Networks of Irish politicians, with color and vertex shapes denoting party affiliation. Politicians in the top qth percentile of* $\sum_k(\Theta + V_m)_{ik}$ *are kept and drawn in the same position as in Figure 1.*

Table 2 in the supplemental article [Mankad and Michailidis (2015)] shows the Structured Semi-NMF measure is again a statistically significant predictor for headline coverage rate, after controlling for all other variables, and Figure 4 shows again that the proposed approach results in an influence measure that improves forecasting accuracy relative to alternative model specifications.

5. Conclusion. The Structured Semi-NMF performs best in both data sets, though the improvement was only slight in the Irish context. The overall results were driven by utilizing all three types of networks for measuring importance and specifying the S_m matrices to boost important politicians with particular types of linkages.

One potential issue with the analysis is that Lexis–Nexis coverage of non-US media and, in particular, the Irish media appears to be imperfect. However, even with poor coverage, as long as it is representative of the overall media landscape, then the reported results will be meaningful. We are also unaware of other tools that can be used for such searches. Another issue is that politicians may appear in headlines that reference their office, for example, "the president." A more comprehensive newspaper headline count is difficult to ascertain, but could in future work provide further validation of the results presented here.

Given that both data sets are exclusively link meta-data, our findings support the notion that the significant challenges associated with content analysis can often be complimented or avoided with network analysis tools for tasks like identifying individuals influential within social networking platforms. We believe this is partly explained by the restriction of the population to politicians and closely related organizations, which ensures to some extent that the unobserved content is both homogeneous and relevant.

A related problem of identifying emergence of key individuals, communities or trends based on network data requires data collected over time. Smoothing strategies, such as in Mankad and Michailidis (2013b), should be useful to extend the given model for network time-series. We believe the proposed model can be useful for applications in marketing and e-commerce, where data is collected on ecosystems that are close to a steady state. Otherwise, as we saw with David Cameron, the model can mischaracterize the importance of key individuals. Specific questions relating to path properties, such as information diffusion [Romero, Meeder and Kleinberg (2011)] or the spread of epidemics [Chew and Eysenbach (2010)], likely require additional methods and techniques specific to those subtopics.

There also has been recent work on a related problem when node features are measured along with network data [Fosdick and Hoff (2013, 2014), Yang, McAuley and Leskovec (2013)]. For instance, one may have access to demographic information or topics and themes of each account's tweets as in Greene, O'Callaghan and Cunningham (2012). While it appears the proposed model could be useful in this setting, using external covariates on the nodes to construct S_m likely raises additional issues that require care, such as variables being available for some, but not all nodes. In this work, the node-level statistics are "internally" calculated directly from the network and, thus, will always cover the full network.

A strength of the Structured Semi-NMF model is that it encompasses different types of links (weighted and binary), integrates information from multiple networks and allows the analyst to utilize contextual knowledge about the given networked system. The method depends upon the analyst choosing appropriate, context-specific node-level statistics. As such, the alternating least squares algorithm provides opportunities for additional regularization in situations where the S_m matrices are high dimensional or when there are no node-specific values that are obvious to use.

SUPPLEMENTARY MATERIAL

Supplement to "Analysis of multiview legislative networks with structured matrix factorization: Does Twitter influence translate to the real world?" (DOI: 10.1214/15-AOAS858SUPP; .pdf). We provide additional simulation results, details and derivations for estimation algorithms, and detailed Poisson regression results.

REFERENCES

Barrat, A., Barthélemy, M., Pastor-Satorras, R. and Vespignani, A. (2004). The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **101** 3747–3752.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. and Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Statist. Data Anal.* **52** 155–173. MR2409971

Brandes, U., Fleischer, D. and Puppe, T. (2006). Dynamic spectral layout of small worlds. In *Graph Drawing* (P. Healy and N. Nikolov, eds.). *Lecture Notes in Computer Science* **3843** 25–36. Springer, Berlin. MR2244497

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, P. K. (2010). Measuring user influence in Twitter: The million follower fallacy. *ICWSM* **10** 10–17.

Chew, C. and Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS ONE* **5** e14118.

Coakley, J. and Gallagher, M. (2005). *Politics in the Republic of Ireland*. Psychology Press, New York.

Ding, C., Li, T. and Jordan, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell.* **32** 45–55.

Fienberg, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *J. Comput. Graph. Statist.* **21** 825–839. MR3005799

Fosdick, B. K. and Hoff, P. D. (2013). Testing and modeling dependencies between a network and nodal attributes. Available at arXiv:1306.4708.

Fosdick, B. K. and Hoff, P. D. (2014). Separable factor analysis with applications to mortality data. *Ann. Appl. Stat.* **8** 120–147. MR3191985

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks* **1** 215–239.

Gemulla, R., Nijkamp, E., Haas, P. J. and Sismanis, Y. (2011). Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 69–77. ACM, New York.

Gillis, N. and Glineur, F. (2008). Nonnegative factorization and the maximum edge biclique problem. Available at arXiv:0810.4225.

Golbeck, J., Grimes, J. M. and Rogers, A. (2010). Twitter use by the U.S. Congress. *J. Am. Soc. Inf. Sci. Technol.* **61** 1612–1621.

Greene, D. and Cunningham, P. (2013). Producing a unified graph representation from multiple social network views. Available at arXiv:1301.5809.

Greene, D., O'Callaghan, D. and Cunningham, P. (2012). Identifying topical Twitter communities via user list aggregation. In *2nd International Workshop on Mining Communities and People Recommenders (COMMPER 2012) at ECML 2012*. Bristol, UK.

Huberman, B. A., Romero, D. M. and Wu, F. (2008). Social networks that matter: Twitter under the microscope. *CoRR* **abs/0812.1045**.

Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer, New York. MR0841268

Katayama, J., Takahashi, N. and Takeuchi, J. (2013). Boundedness of modified multiplicative updates for nonnegative matrix factorization. In *IEEE 5th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)* 252–255. St. Martin.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM* **46** 604–632. MR1747649

Koren, Y. (2005). Drawing graphs by eigenvectors: Theory and practice. *Comput. Math. Appl.* **49** 1867–1888. MR2154691

KROONENBERG, P. M. and DE LEEUW, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika* **45** 69–97. MR0570771

LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401** 788–791.

LIN, Y.-R., CHI, Y., ZHU, S., SUNDARAM, H. and TSENG, B. L. (2008). Facetnet: A framework for analyzing communities and their evolutions in dynamic networks. In *Proceedings of the 17th International Conference on World Wide Web*. 685–694. ACM, New York.

MANKAD, S. and MICHAILIDIS, G. (2013a). Discovery of path-important nodes using structured semi-nonnegative matrix factorization. In *IEEE 5th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)* 288–291. St. Martin.

MANKAD, S. and MICHAILIDIS, G. (2013b). Structural and functional discovery in dynamic networks with non-negative matrix factorization. *Phys. Rev. E* **88** 042812.

MANKAD, S. and MICHAILIDIS, G. (2015). Supplement to "Analysis of multiview legislative networks with structured matrix factorization: Does Twitter influence translate to the real world?" DOI:10.1214/15-AOAS858SUPP.

MCKELVEY, K., DIGRAZIA, J. and ROJAS, F. (2014). Twitter publics: How online political communities signaled electoral outcomes in the 2010 US house election. *Information, Communication & Society* **17** 436–450.

NEWMAN, M. E. J. (2010). *Networks*. Oxford Univ. Press, Oxford. MR2676073

OWEN, A. B. and PERRY, P. O. (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Ann. Appl. Stat.* **3** 564–594. MR2750673

PAGE, L., BRIN, S., MOTWANI, R. and WINOGRAD, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, Stanford, CA. Available at: `http://ilpubs.stanford.edu:8090/422/`.

PSORAKIS, I., ROBERTS, S., EBDEN, M. and SHELDON, B. (2011). Overlapping community detection using Bayesian non-negative matrix factorization. *Phys. Rev. E* **83** 066114.

RECHT, B. and RÉ, C. (2013). Parallel stochastic gradient algorithms for large-scale matrix completion. *Math. Program. Comput.* **5** 201–226. MR3069879

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856

ROHE, K. and YU, B. (2012). Co-clustering for directed graphs; the stochastic co-blockmodel and a spectral algorithm. Available at arXiv:1204.2296.

ROMERO, D. M., MEEDER, B. and KLEINBERG, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*. 695–704. ACM, New York.

SALTER-TOWNSHEND, M. and MURPHY, T. B. (2015). Role analysis in networks using mixtures of exponential random graph models. *J. Comput. Graph. Statist.* **24** 520–538.

SALTER-TOWNSHEND, M., WHITE, A., GOLLINI, I. and MURPHY, T. B. (2012). Review of statistical network analysis: Models, algorithms, and software. *Stat. Anal. Data Min.* **5** 260–264. MR2958152

THE NEW YORK TIMES BLOGS (2011). Twitter Starts Selling Political Ads. Available at `http://thecaucus.blogs.nytimes.com/2011/09/21/twitter-starts-selling-political-ads/`. Accessed: 2013-11-13.

THE NEW YORK TIMES BLOGS (2012). Pepsi and Twitter Announce Partnership on Ad Campaign. Available at `http://mediadecoder.blogs.nytimes.com/2012/05/30/pepsi-and-twitter-announce-partnership-on-ad-campaign`. Accessed: 2013-11-13.

The New York Times (2013). Using Twitter to Move the Markets. http://www.nytimes.com/2013/10/07/business/media/using-twitter-to-move-the-markets.html. Accessed: 2013-11-13.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. and Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* 178–185. Washington, DC.

Twitter, Inc. (2014). About Twitter, Inc. Available at https://about.twitter.com/company. Accessed: 2014-09-19.

Unankard, S., Li, X., Sharaf, M., Zhong, J. and Li, X. (2014). Predicting elections from social networks based on sub-event detection and sentiment analysis. In *Web Information Systems Engineering—WISE 2014* (B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakali and Y. Zhang, eds.). *Lecture Notes in Computer Science* **8787** 1–16. Springer, Berlin.

Wang, F., Li, T., Wang, X., Zhu, S. and Ding, C. (2011). Community discovery using nonnegative matrix factorization. *Data Min. Knowl. Discov.* **22** 493–521. MR2785131

Xu, W., Liu, X. and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* 267–273. ACM, New York.

Yang, J., McAuley, J. and Leskovec, J. (2013). Community detection in networks with node attributes. In *IEEE 13th International Conference on Data Mining (ICDM)* 1151–1156. IEEE, New York.

Operations, Technology and Information Management
Cornell University
Ithaca, New York 14850
USA
E-mail: smankad@cornell.edu

Department of Statistics
University of Michigan
Ann Arbor, Michigan 48109
USA
E-mail: gmichail@umich.edu